Análisis Inteligente de Datos: Segundo Parcial

Claudio Sebastián Castillo

16 de mayo de 2022

Contents

1 Pregunta1

1.1 EDA

1.1.1 structure

```
spec_tbl_df [5 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Repeticion: num [1:5] 1 2 3 4 5
 $ Lote1     : num [1:5] 29.4 31.5 30.9 27.6 28.9
 $ Lote2     : num [1:5] 30.6 32.1 30.1 29.6 29.7
 $ Lote3     : num [1:5] 27.2 26.6 25.3 27.7 27.1
 $ Lote4     : num [1:5] 31 31 28.9 31.4 29.7
 $ Lote5     : num [1:5] 29.7 29.3 26.9 31.6 29.4
 - attr(*, "spec")=
  .. cols(
```

```
  ..   Repeticion = col_double(),
  ..   Lote1 = col_double(),
  ..   Lote2 = col_double(),
  ..   Lote3 = col_double(),
  ..   Lote4 = col_double(),
  ..   Lote5 = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

1.1.2 Summary

```
   Repeticion     Lote1           Lote2           Lote3           Lote4      
 Min.   :1    Min.   :27.63   Min.   :29.63   Min.   :25.31   Min.   :28.95  
 1st Qu.:2    1st Qu.:28.85   1st Qu.:29.68   1st Qu.:26.63   1st Qu.:29.70  
 Median :3    Median :29.39   Median :30.11   Median :27.10   Median :30.98  
 Mean   :3    Mean   :29.65   Mean   :30.43   Mean   :26.77   Mean   :30.42  
 3rd Qu.:4    3rd Qu.:30.88   3rd Qu.:30.63   3rd Qu.:27.16   3rd Qu.:31.03  
 Max.   :5    Max.   :31.51   Max.   :32.10   Max.   :27.66   Max.   :31.45  
     Lote5      
 Min.   :26.87  
 1st Qu.:29.32  
 Median :29.41  
 Mean   :29.37  
 3rd Qu.:29.67  
 Max.   :31.59  
```

1.1.3 Control NAs

```
# A tibble: 1 x 6
  Repeticion Lote1 Lote2 Lote3 Lote4 Lote5
       <int> <int> <int> <int> <int> <int>
1          0     0     0     0     0     0
```

1.1.4 Distribución de datos

```
$coeficiente_variacion
# A tibble: 1 x 6
  Repeticion Lote1 Lote2 Lote3 Lote4 Lote5
       <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1       52.7  5.27  3.34  3.34  3.46  5.72

$sesgo
# A tibble: 1 x 6
  Repeticion   Lote1 Lote2  Lote3  Lote4  Lote5
       <dbl>   <dbl> <dbl>  <dbl>  <dbl>  <dbl>
1          0 -0.0392 0.984 -0.910 -0.500 -0.279

$curtosis
# A tibble: 1 x 6
  Repeticion Lote1 Lote2 Lote3 Lote4 Lote5
       <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1        1.7  1.64  2.49  2.54  1.61  2.49

$mad
# A tibble: 1 x 6
```

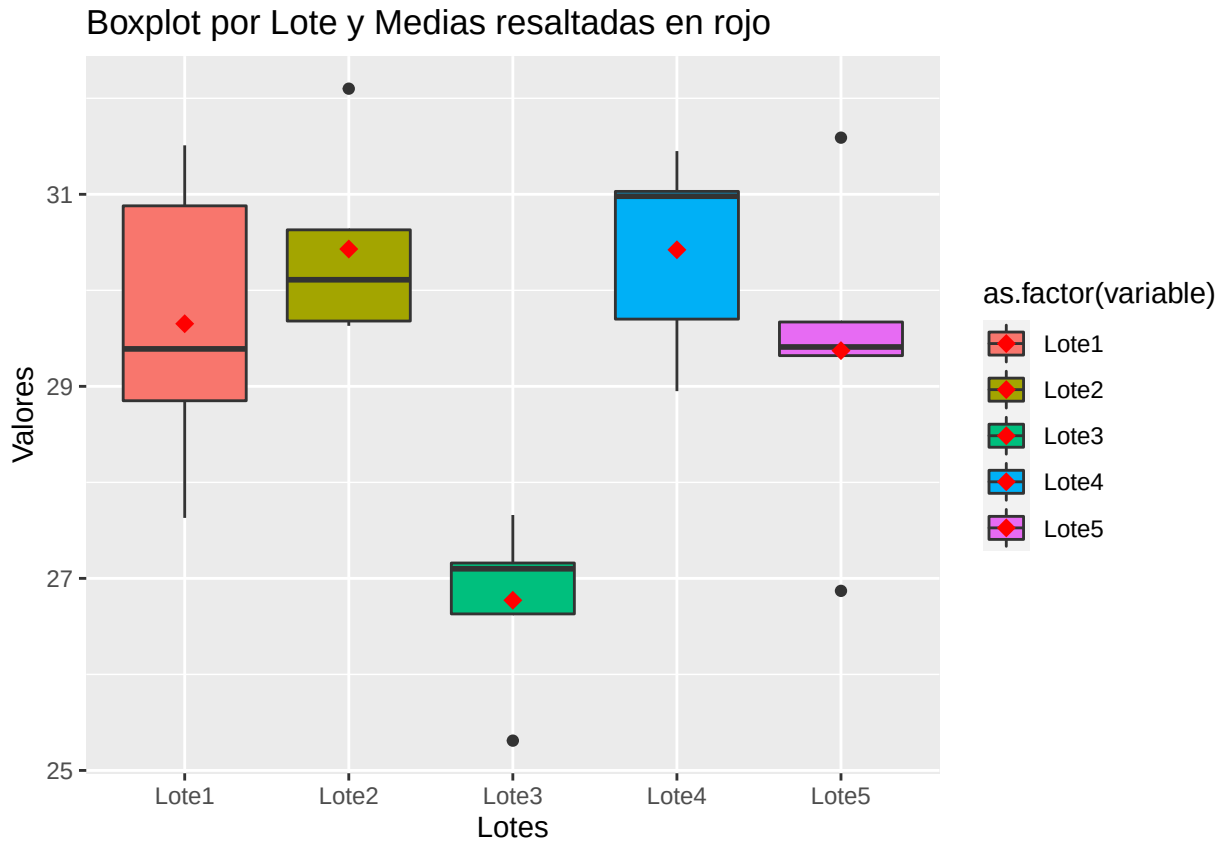
```
  Repeticion Lote1 Lote2 Lote3 Lote4 Lote5
       <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1       1.48  2.21 0.712 0.697 0.697 0.385
```

```
$m_correlacion
           Repeticion Lote1 Lote2 Lote3 Lote4 Lote5
Repeticion       1.00 -0.50 -0.68  0.16 -0.33  0.16
Lote1           -0.50  1.00  0.77 -0.75 -0.37 -0.74
Lote2           -0.68  0.77  1.00 -0.18  0.29 -0.15
Lote3            0.16 -0.75 -0.18  1.00  0.79  0.96
Lote4           -0.33 -0.37  0.29  0.79  1.00  0.86
Lote5            0.16 -0.74 -0.15  0.96  0.86  1.00
```

1.1.5 Vector de Medias

```
 Lote1  Lote2  Lote3  Lote4  Lote5
29.652 30.430 26.772 30.422 29.372
```

1.1.6 Datos sobre valores por Lote



1.1.7 Contraste entre el lote 2 y lote 3 mediante t-test

```
    Two Sample t-test

data:  temp$Lote2 and temp$Lote3
t = 6.039, df = 8, p-value = 0.0003097
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 2.261193 5.054807
sample estimates:
mean of x mean of y
   30.430    26.772
```

```
[1] "H0 debe rechazarse, los grupos son distintos a nivel de significancia 0.05"
```

1.2 Respuestas

Presupusto del Estudio: las puntuaciones bajas se consideran baja resistencia a la torsion.

1) Hipótesis acerca de la resistencia de materiales: -Hipótesis nula: todos los proveedores ofrecen materiales cuya resistencia media son iguales entre sí. -Hipótesis alternativa: al menos dos medias son distintas.

2) Análisis: El data set no contiene errores de registro ni datos faltantes, y en tal sentido es propiado para ser evaluado. Se evidencian valores extremos en los datos de resistencia en los Lotes 2, 3, y 5, que dado los pocos datos disponibles no se excluirán del estudio. Considerando los valores medios de la resistencia de los materiales de cada lote surge que el proveedor con datos de menor resistencia promedio en los materiales es el Lote 3 con un valor promedio de 26.772, mientras que aquellos de mayor resistencia promedio son el lote 2 con valores de 30.430 y el lote 4 con 30.422. Se advierte también que la diferencia en los promedios de resistencia entre el lote 2 y lote 4 es poco significativa (0.008 ptos.), pero a favor del lote 2 tiene una menor variabilidad en sus datos.

Se efectuó un T-test para contraste entre las medias de los lotes 2 y 3 para determinar si sus diferencias eran significativas, y se enoncotró evidencia favorable en tal sentido (con un p-value = 0.0003097) con lo que rechazamos la hipótesis de medias iguales entre ambos grupos.

3) Conclusión:

- El lote con peor promedio de resistencia en sus materiales es el lote 3, por lo que NO RECOMENDAMOS su inclusión consideración para futuras compras.

- El lote con mejores resultados generales según lo expuesto en el análisis y recomendado para compras futuras es el lote 2.

2 Pregunta2

2.1 EDA

2.1.1 structure

```
tibble [157 x 14] (S3: tbl_df/tbl/data.frame)
 $ marca    : chr [1:157] "Acura" "Acura" "Acura" "Acura" ...
 $ modelo   : chr [1:157] "Integra" "TL" "CL" "RL" ...
 $ venta    : num [1:157] 16.92 39.38 14.11 8.59 20.4 ...
 $ reventa  : num [1:157] 16.4 19.9 18.2 29.7 22.3 ...
 $ tipo     : num [1:157] 0 0 0 0 0 0 0 0 0 0 ...
 $ precio   : num [1:157] 21.5 28.4 NA 42 24 ...
 $ motor    : num [1:157] 1.8 3.2 3.2 3.5 1.8 2.8 4.2 2.5 2.8 2.8 ...
 $ CV       : num [1:157] 140 225 225 210 150 200 310 170 193 193 ...
 $ pisada   : num [1:157] 101 108 107 115 103 ...
 $ ancho    : num [1:157] 67.3 70.3 70.6 71.4 68.2 76.1 74 68.4 68.5 70.9 ...
 $ largo    : num [1:157] 172 193 192 197 178 ...
 $ peso     : num [1:157] 2.64 3.52 3.47 3.85 3 ...
 $ depóstito: num [1:157] 13.2 17.2 17.2 18 16.4 18.5 23.7 16.6 16.6 18.5 ...
 $ mpg      : num [1:157] 28 25 26 22 27 22 21 26 24 25 ...
```

2.1.2 Summary

```
   marca              modelo              venta           reventa      
 Length:157         Length:157         Min.   :  0.11   Min.   : 5.16  
 Class :character   Class :character   1st Qu.: 14.11   1st Qu.:11.26  
 Mode  :character   Mode  :character   Median : 29.45   Median :14.18  
                                       Mean   : 53.00   Mean   :18.07  
                                       3rd Qu.: 67.96   3rd Qu.:19.88  
                                       Max.   :540.56   Max.   :67.55  
                                                        NA's   :36     
     tipo            precio           motor             CV       
 Min.   :0.0000   Min.   : 9.235   Min.   :1.000   Min.   : 55.0  
 1st Qu.:0.0000   1st Qu.:18.017   1st Qu.:2.300   1st Qu.:149.5  
 Median :0.0000   Median :22.799   Median :3.000   Median :177.5  
 Mean   :0.2611   Mean   :27.391   Mean   :3.061   Mean   :185.9  
 3rd Qu.:1.0000   3rd Qu.:31.948   3rd Qu.:3.575   3rd Qu.:215.0  
 Max.   :1.0000   Max.   :85.500   Max.   :8.000   Max.   :450.0  
                  NA's   :2        NA's   :1       NA's   :1      
    pisada          ancho           largo            peso      
 Min.   : 92.6   Min.   :62.60   Min.   :149.4   Min.   :1.895  
 1st Qu.:103.0   1st Qu.:68.40   1st Qu.:177.6   1st Qu.:2.971  
 Median :107.0   Median :70.55   Median :187.9   Median :3.342  
 Mean   :107.5   Mean   :71.15   Mean   :187.3   Mean   :3.378  
 3rd Qu.:112.2   3rd Qu.:73.42   3rd Qu.:196.1   3rd Qu.:3.800  
 Max.   :138.7   Max.   :79.90   Max.   :224.5   Max.   :5.572  
 NA's   :1       NA's   :1       NA's   :1       NA's   :2      
  depóstito          mpg       
 Min.   :10.30   Min.   :15.00  
 1st Qu.:15.80   1st Qu.:21.00  
 Median :17.20   Median :24.00  
 Mean   :17.95   Mean   :23.84  
 3rd Qu.:19.57   3rd Qu.:26.00  
 Max.   :32.00   Max.   :45.00  
```

```
 NA's   :1       NA's   :3
```

2.1.3 Control NAs

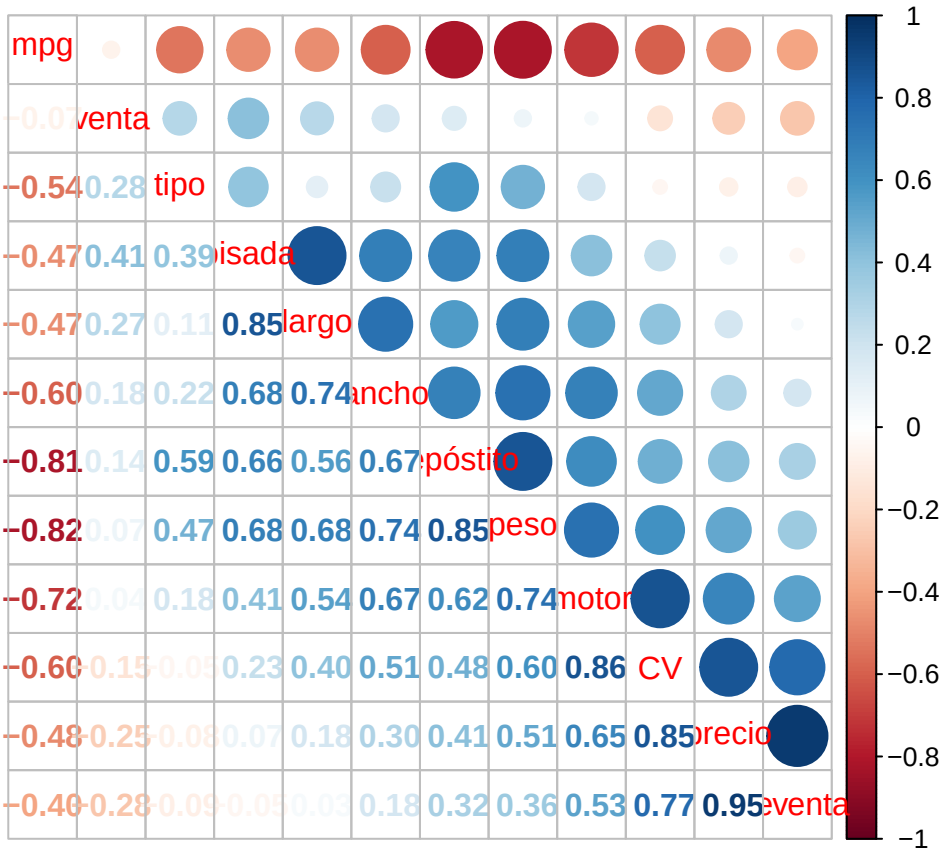
```
# A tibble: 1 x 14
  marca modelo venta reventa  tipo precio motor    CV pisada ancho largo  peso
  <int>  <int> <int>   <int> <int>  <int> <int> <int>  <int> <int> <int> <int>
1     0      0     0      36     0      2     1     1      1     1     1     2
# ... with 2 more variables: depóstito <int>, mpg <int>
```

2.1.4 Distribución de datos

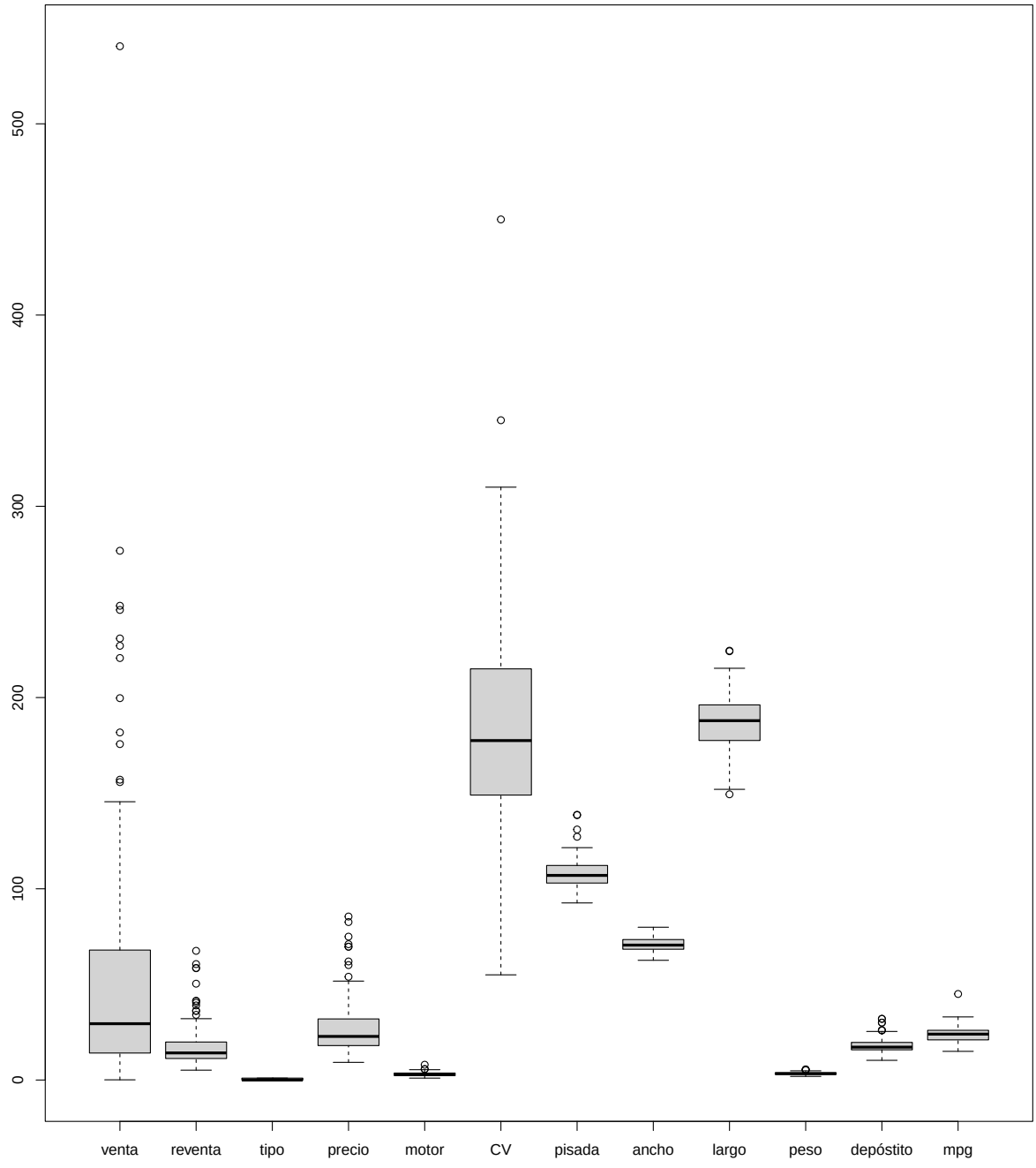
```
$coeficiente_variacion
# A tibble: 1 x 12
  venta reventa  tipo precio motor    CV pisada ancho largo  peso depóstito
  <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>     <dbl>
1  128.    63.4  169.   52.4  34.1  30.5   7.11  4.85  7.17  18.7      21.7
# ... with 1 more variable: mpg <dbl>

$sesgo
# A tibble: 1 x 12
  venta reventa  tipo precio motor    CV pisada ancho largo  peso depóstito
  <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>     <dbl>
1  3.38      NA  1.09     NA    NA    NA     NA    NA    NA    NA        NA
# ... with 1 more variable: mpg <dbl>

$curtosis
# A tibble: 1 x 12
  venta reventa  tipo precio motor    CV pisada ancho largo  peso depóstito
  <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>     <dbl>
1  20.0      NA  2.18     NA    NA    NA     NA    NA    NA    NA        NA
# ... with 1 more variable: mpg <dbl>

$mad
# A tibble: 1 x 12
  venta reventa  tipo precio motor    CV pisada ancho largo  peso depóstito
  <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>     <dbl>
1  30.3      NA     0     NA    NA    NA     NA    NA    NA    NA        NA
# ... with 1 more variable: mpg <dbl>

$m_correlacion
          venta reventa  tipo precio motor    CV pisada ancho largo  peso
venta      1.00   -0.28  0.28  -0.25  0.04 -0.15   0.41  0.18  0.27  0.07
reventa   -0.28    1.00 -0.09   0.95  0.53  0.77  -0.05  0.18  0.03  0.36
tipo       0.28   -0.09  1.00  -0.08  0.18 -0.05   0.39  0.22  0.11  0.47
precio    -0.25    0.95 -0.08   1.00  0.65  0.85   0.07  0.30  0.18  0.51
motor      0.04    0.53  0.18   0.65  1.00  0.86   0.41  0.67  0.54  0.74
CV        -0.15    0.77 -0.05   0.85  0.86  1.00   0.23  0.51  0.40  0.60
pisada     0.41   -0.05  0.39   0.07  0.41  0.23   1.00  0.68  0.85  0.68
ancho      0.18    0.18  0.22   0.30  0.67  0.51   0.68  1.00  0.74  0.74
largo      0.27    0.03  0.11   0.18  0.54  0.40   0.85  0.74  1.00  0.68
peso       0.07    0.36  0.47   0.51  0.74  0.60   0.68  0.74  0.68  1.00
depóstito  0.14    0.32  0.59   0.41  0.62  0.48   0.66  0.67  0.56  0.85
mpg       -0.07   -0.40 -0.54  -0.48 -0.72 -0.60  -0.47 -0.60 -0.47 -0.82
          depóstito   mpg
```

```
venta          0.14 -0.07
reventa        0.32 -0.40
tipo           0.59 -0.54
precio         0.41 -0.48
motor          0.62 -0.72
CV             0.48 -0.60
pisada         0.66 -0.47
ancho          0.67 -0.60
largo          0.56 -0.47
peso           0.85 -0.82
depóstito      1.00 -0.81
mpg           -0.81  1.00
```

2.1.5 Grafico Correlaciones



2.1.6 Boxplot variables numericas



2.2 Observaciones en torno a la exploración de datos

-Observamos la presencia de valore etremos en muchas variables.
-Observamos que las variables estarían bien formadas aunque hay muchos valores faltantes en algunas variables (ej. Reventa tiene 36 NAs) -Observamos también posibles diferencias en las escalas de medición (ej.motor vs

ventas).
-Observamos que hay variables con una fuerte correlación, por lo que podría reducirse la dimensionalidad del problema para mejor la comprensión de los datos.

2.3 Analisis de Autos en base a clustering

En este análisi buscaremos patrones o grupos dentro de un conjunto de observaciones con el fin de encontrar similaridades en los autos y marcas a fin de proponer estrategias para su tratamiento diferencial.

2.3.1 Se centran los datos de la matriz

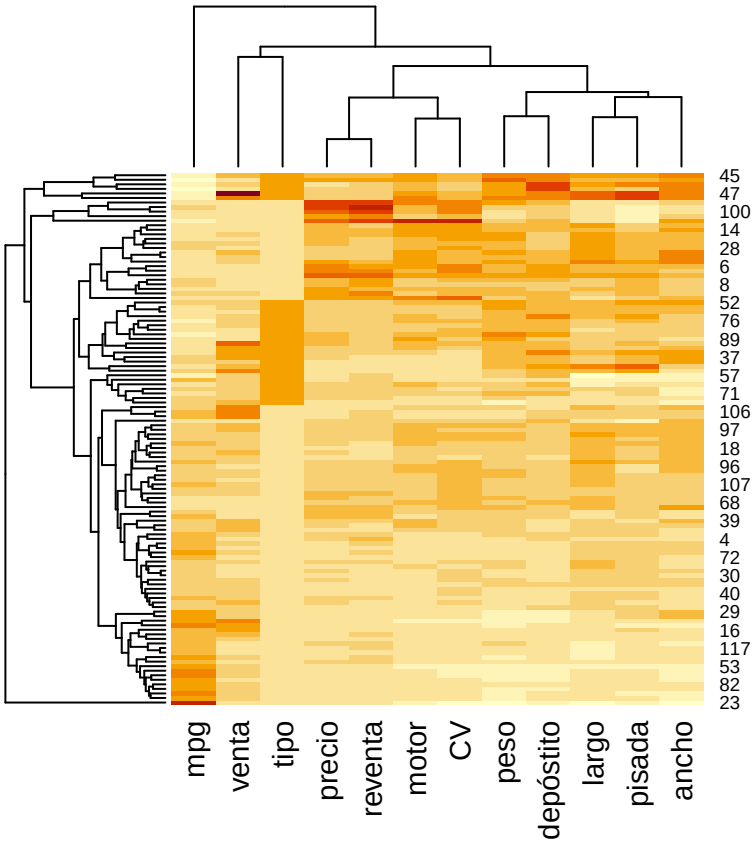
Considerando que las variables estén registradas en diferentes escalas y/o unidades de medición, vamos a centrar las mismas (media 0 y desviación estandar 1) de tal manera de evitar la influencia de las unidades de medición en la clusterización.

Se omiten datos faltantes.

2.3.2 Se calculas distancias euclideas

```
     1    2    3    4    5
1 0.00 3.42 4.71 1.31 4.29
2 3.42 0.00 1.90 2.57 1.99
3 4.71 1.90 0.00 3.75 1.94
4 1.31 2.57 3.75 0.00 3.35
5 4.29 1.99 1.94 3.35 0.00
```

2.4 Heatmap con Cluster Laterales



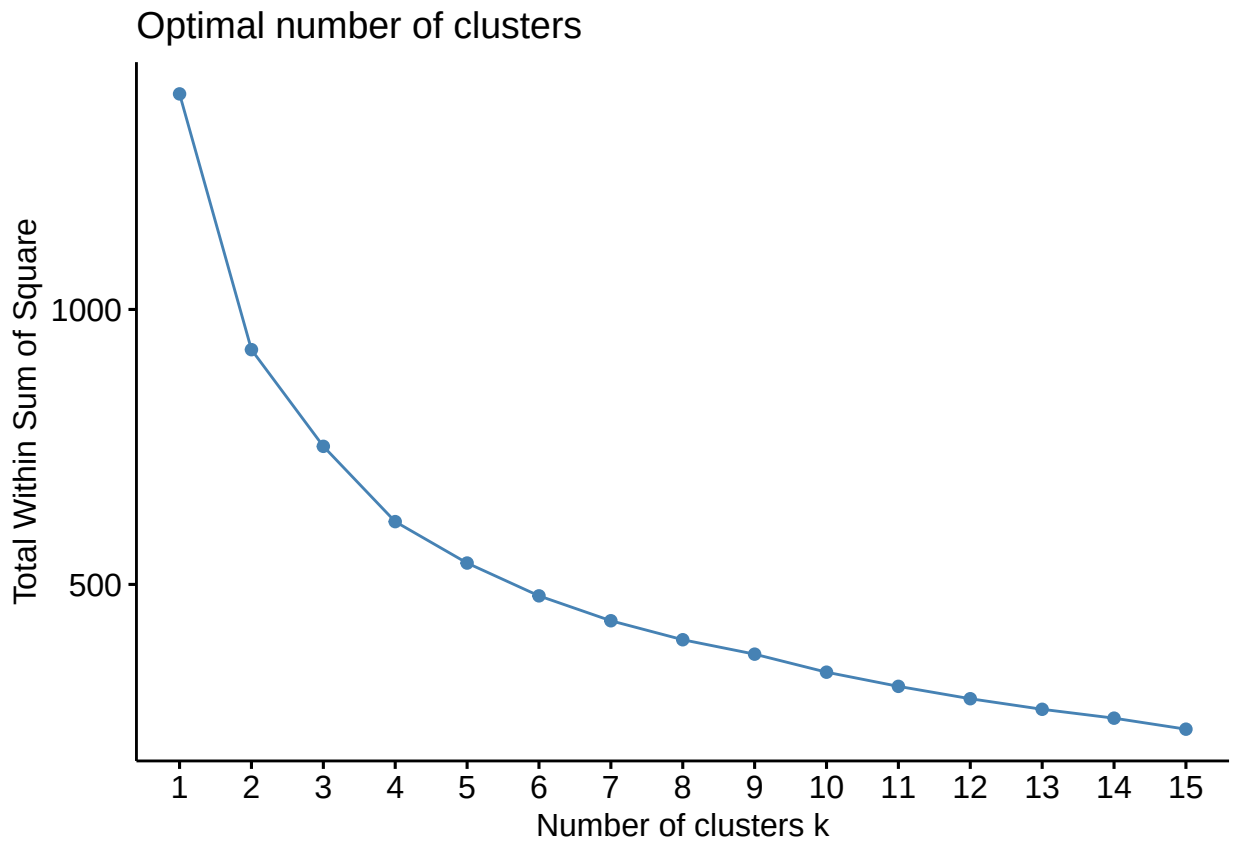
A través de este heatmap comenzamos a aprecier grupos que surgen considerando el color. Así, el color rojo representa alta similaridad y el color azul baja similaridad.

Por lo visto an los presentaciones anteriores procederemos a aplicar distintas técnicas de agrupamiento.

2.5 Cluster No Jerarquico: K-means (x centroides)

2.5.1 Seleccion de los aglomerados de autos (base: método elbow)

Una forma de estimar el número agplomerados o cluster óptimo a falta de informacion adicional, es aplicar el algoritmo de K-means e identificar aquel valor a partir del cual la reducción en la suma total de varianza intra-cluster deja de ser sustancial. A esta estrategia se la conoce como método del codo o elbow method.



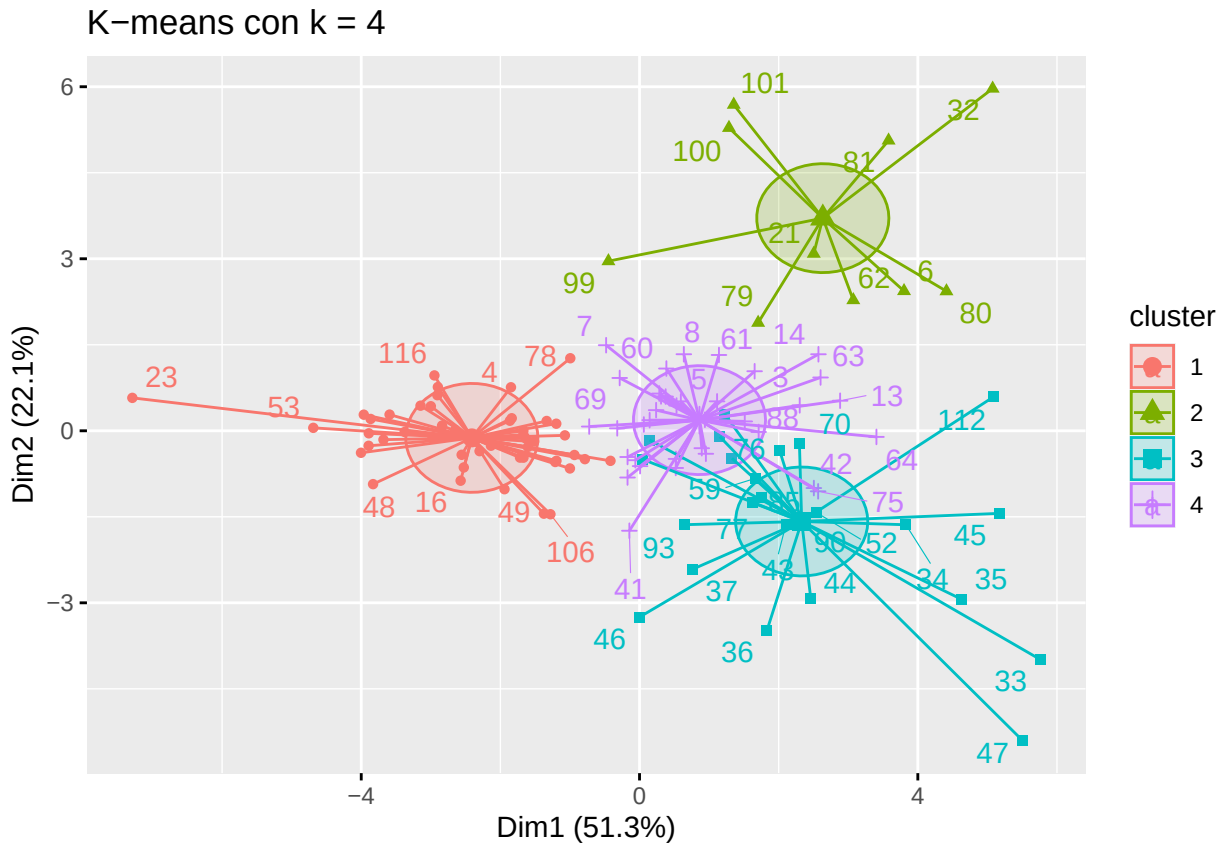
Advertimos que 4 grupos es un númoero optimo.

2.5.2 Resultado

```
K-means clustering with 4 clusters of sizes 47, 10, 24, 36

Cluster means:
         venta     reventa       tipo      precio      motor          CV
1 -0.001223397 -0.53442927 -0.3262702 -0.66461663 -0.8648355 -0.81060539
2 -0.672391867  2.70217607 -0.5716020  2.55383608  1.2711536  2.04325482
3  0.578278608 -0.13237798  1.7345164 -0.02996145  0.5264073  0.06132171
4 -0.197146341  0.03537462 -0.5716020  0.17826932  0.4250544  0.44983844
      pisada      ancho      largo       peso   depóstito        mpg
1 -0.5944995 -0.7890565 -0.7205563 -0.9289958 -0.80182376  0.7747259
2 -0.4343156  0.2181936 -0.2619472  0.3520712  0.59483847 -0.6401811
```

```
3  0.9961183  0.7063692  0.4803552  1.0998816  1.29451528 -1.0677770
4  0.2327165  0.4986349  0.6932525  0.3818037  0.01858237 -0.1217683

Clustering vector:
  [1] 1 4 4 1 4 2 4 4 4 4 4 4 4 4 4 4 1 1 4 4 4 2 1 1 1 4 4 1 4 1 1 1 2 3 3 3 3 3
 [38] 1 4 1 4 4 3 3 3 3 3 1 1 1 3 3 1 1 1 4 1 3 3 4 4 2 4 4 1 1 1 4 4 3 3 1 1 4
 [75] 4 3 3 1 2 2 2 1 1 4 3 3 1 4 3 3 1 1 3 1 1 4 4 4 2 2 2 1 1 1 1 1 4 1 1 1 1
[112] 3 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 192.0041  84.3612 226.8326 110.9619
 (between_SS / total_SS =  55.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

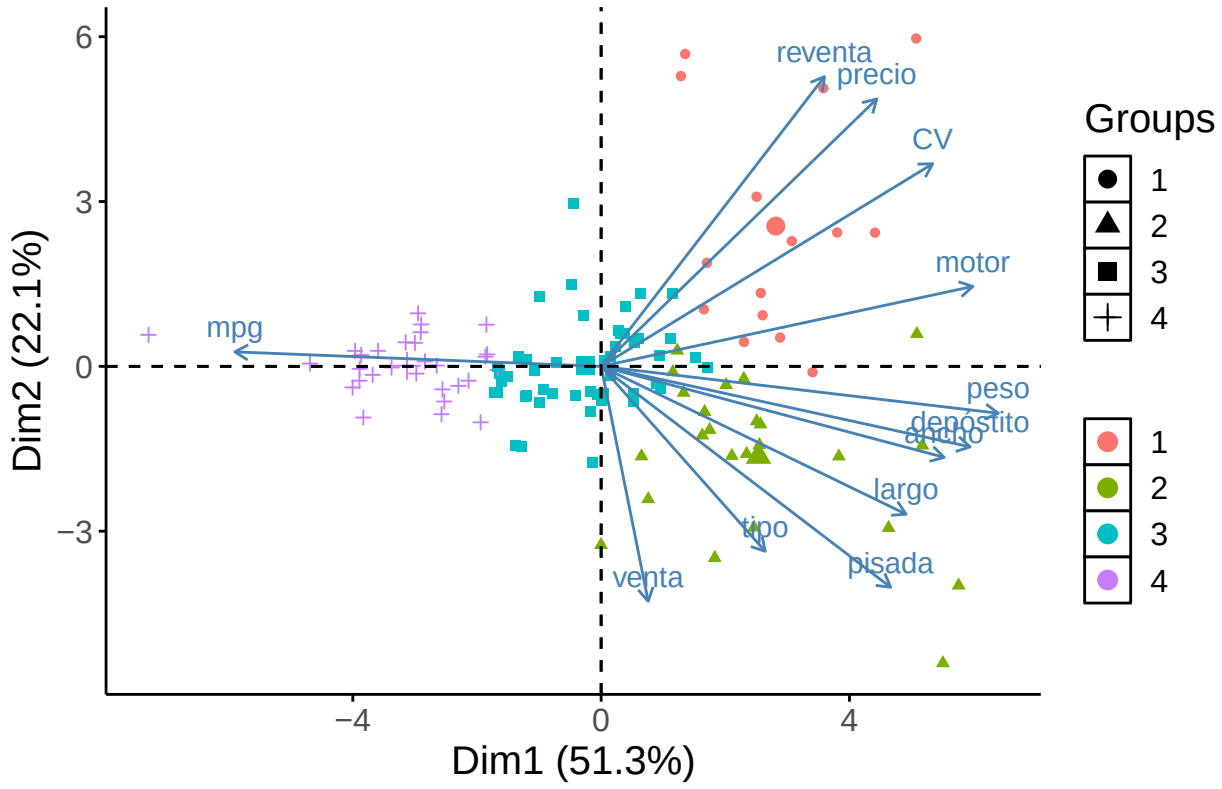
Esta primera aproximación nos arroja la siguiente representación.



Vemos que hay un agrupamiento aceptable de las observaciones. Sin perjuicio de ello se encuentra solapamiento entre los cluster 3 y 4.

2.5.3 Grafico Cluster con PCA

Considerando la dimensión del data set, a continuación reduciremos su magnitud mediante el análisis de componentes principales.

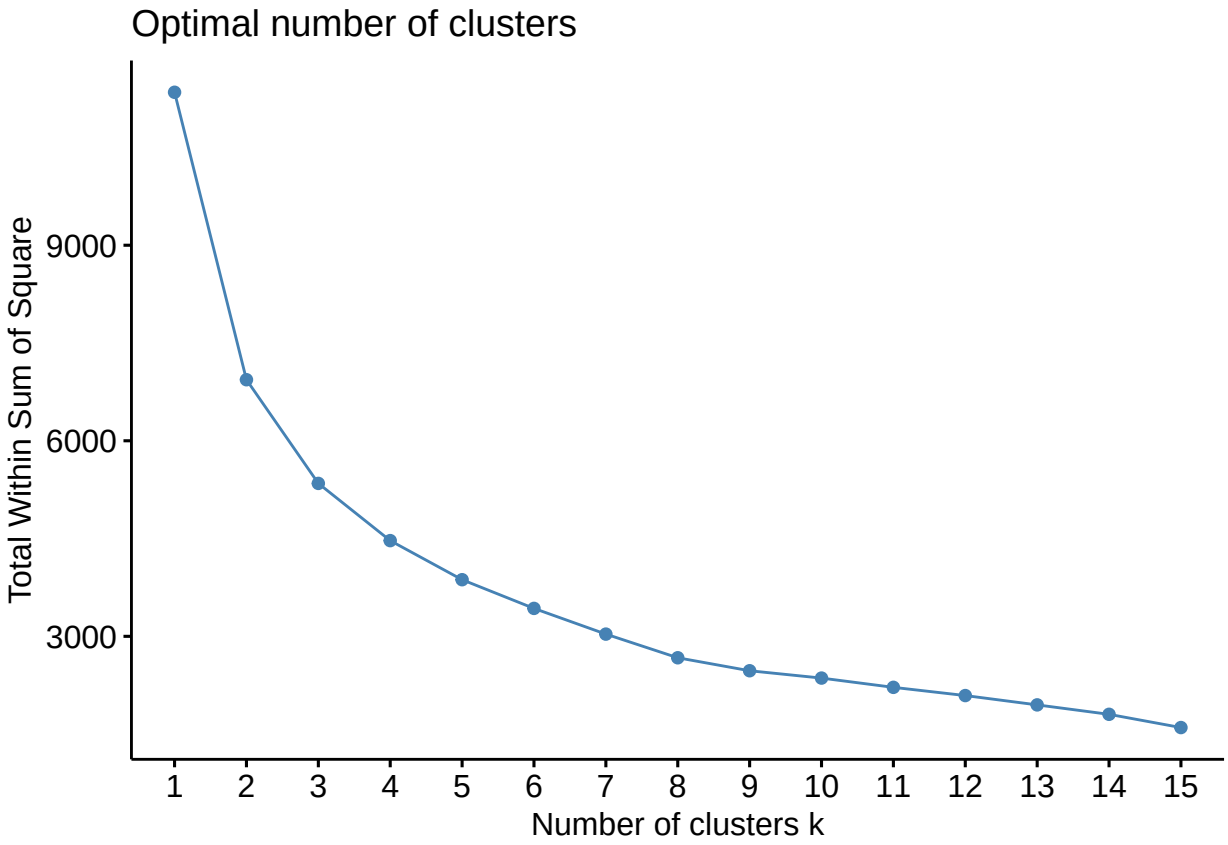


Vemos que mpg tiene un peso predominante en a primera componente, y ventas en la componente 2. Asimismo vemos el contraste entre las variable de consumo y el resto, como tambièn la sepración entre los grupos.

Para proseguir realizaremos nuevos agrupamientos intentando establecer orden de jerarquía.

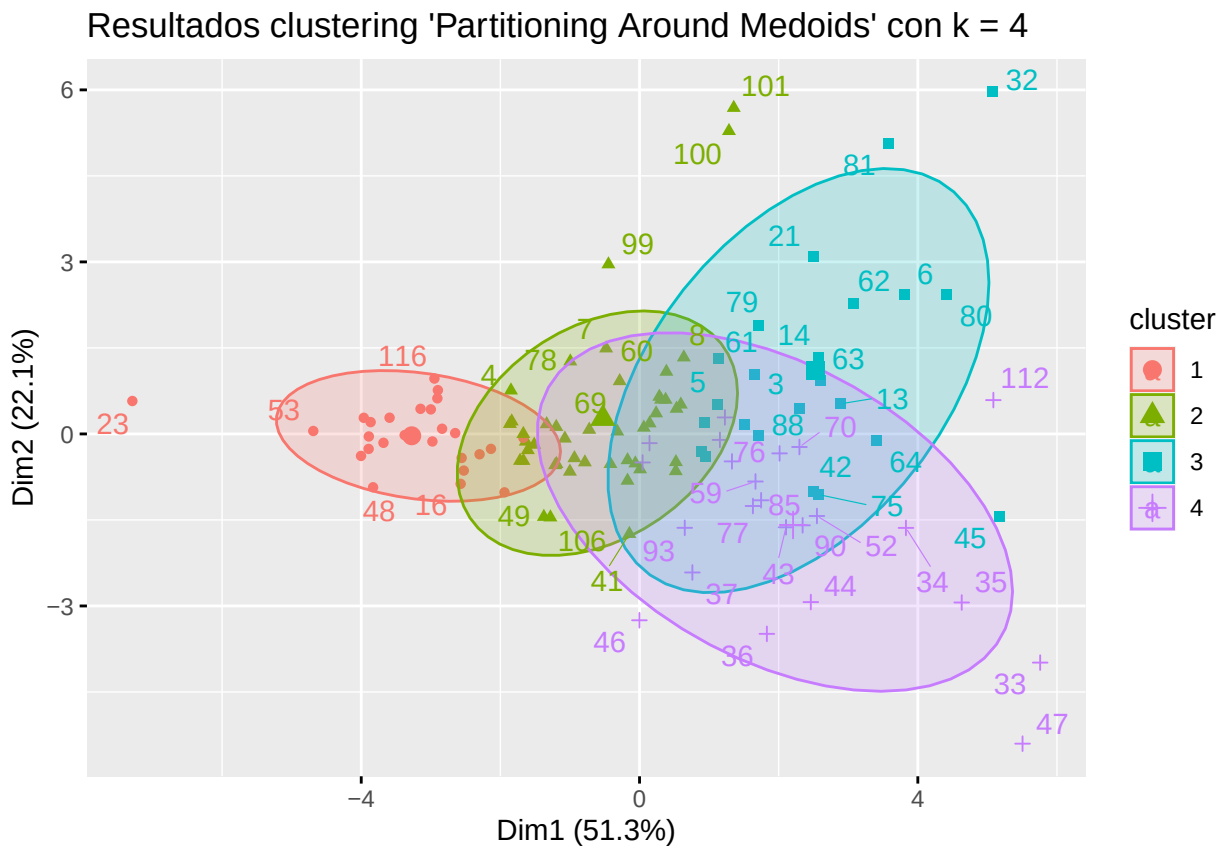
2.6 Cluster No Jerarquico: K-medoids clustering (con centro en observación más representativa)

2.6.1 Selección de k con distancia de Manhattan como medida de similitud




```
Medoids:
     ID      venta    reventa      tipo      precio      motor         CV
[1,] 82 -0.2194185 -0.8255938 -0.571602 -0.88132524 -1.1834292 -0.9435120
[2,] 25 -0.3508885 -0.3318680 -0.571602 -0.10420625 -0.5200284 -0.2266879
[3,] 88 -0.5918325  0.1601345 -0.571602  0.72506931  0.9015446  1.1728256
[4,] 85 -0.4237246 -0.2284700  1.734516  0.03035491  0.2381439 -0.1925535
         pisada      ancho      largo       peso  depóstito         mpg
[1,] -0.9349001 -1.1018634 -0.7377620 -1.2241791 -1.2156247  1.33508516
[2,] -0.1647700 -0.5636427  0.3813776  0.0133105 -0.4777358 -0.02716743
[3,]  0.8041033  0.9093823  1.2766892  1.0766473  0.1810936 -0.48125163
[4,]  0.6053601  1.0510193  0.5113422  1.1168364  0.5763912 -0.70829373
Clustering vector:
  [1] 1 2 3 2 3 3 2 2 2 3 3 3 3 3 2 1 2 2 2 2 3 1 1 2 2 2 2 3 1 2 2 3 4 4 4 4 4
 [38] 1 2 2 2 3 4 4 3 4 4 1 2 2 4 4 1 1 2 2 1 4 4 2 3 3 3 3 1 2 2 2 2 4 4 1 1 2
 [75] 3 4 4 2 3 3 3 1 2 2 4 4 2 3 4 4 1 2 4 2 2 2 2 3 2 2 2 1 1 1 1 2 2 1 1 1 2
[112] 4 1 1 2 1 1
Objective function:
   build     swap
5.600272 5.600272

Available components:
 [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
```

```
[6] "clusinfo"    "silinfo"     "diss"        "call"        "data"
```



2.7 Cluster Jerárquicos

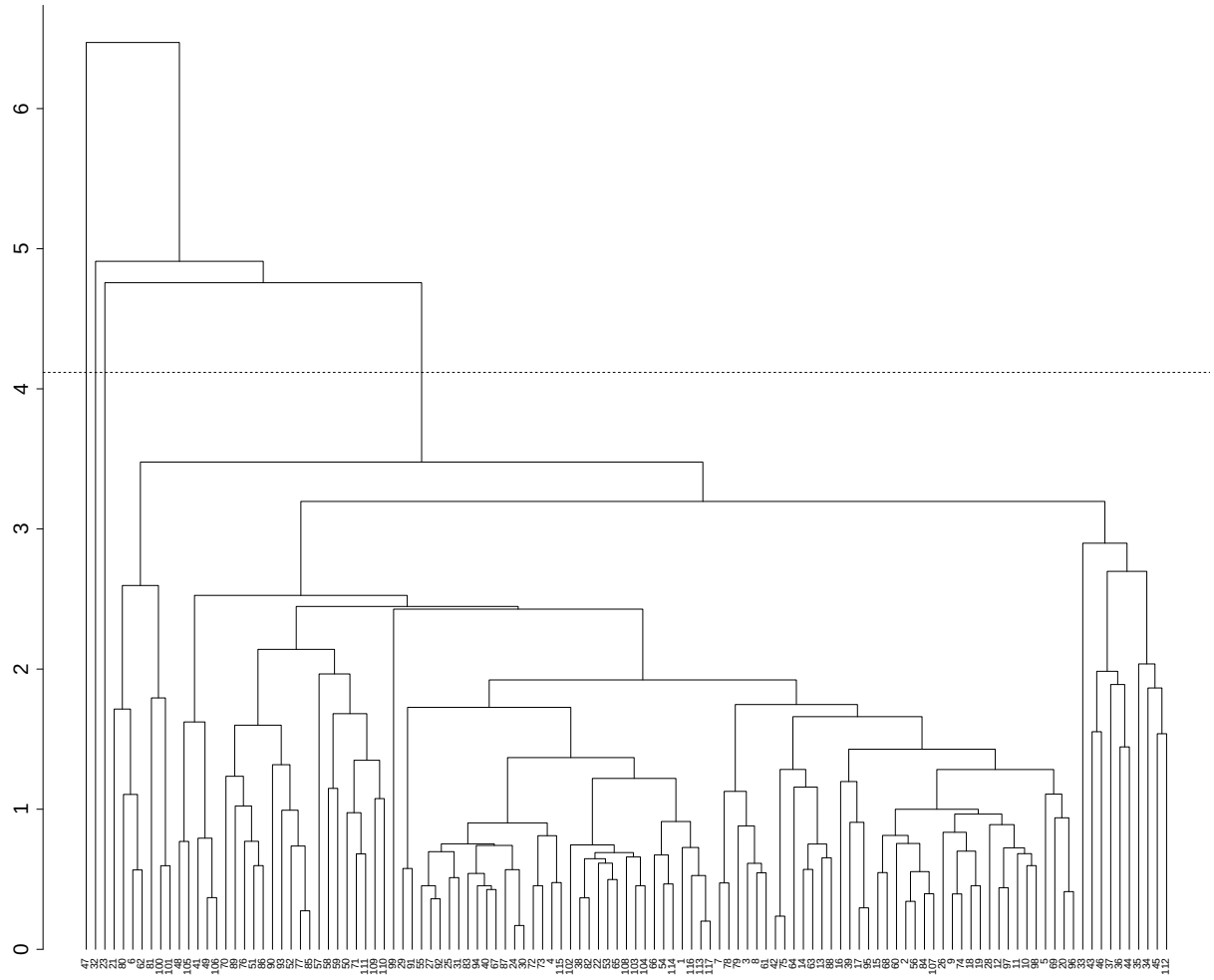
2.8 Modelo óptimo considerando distintas matrices de distancias y linkage intercluster

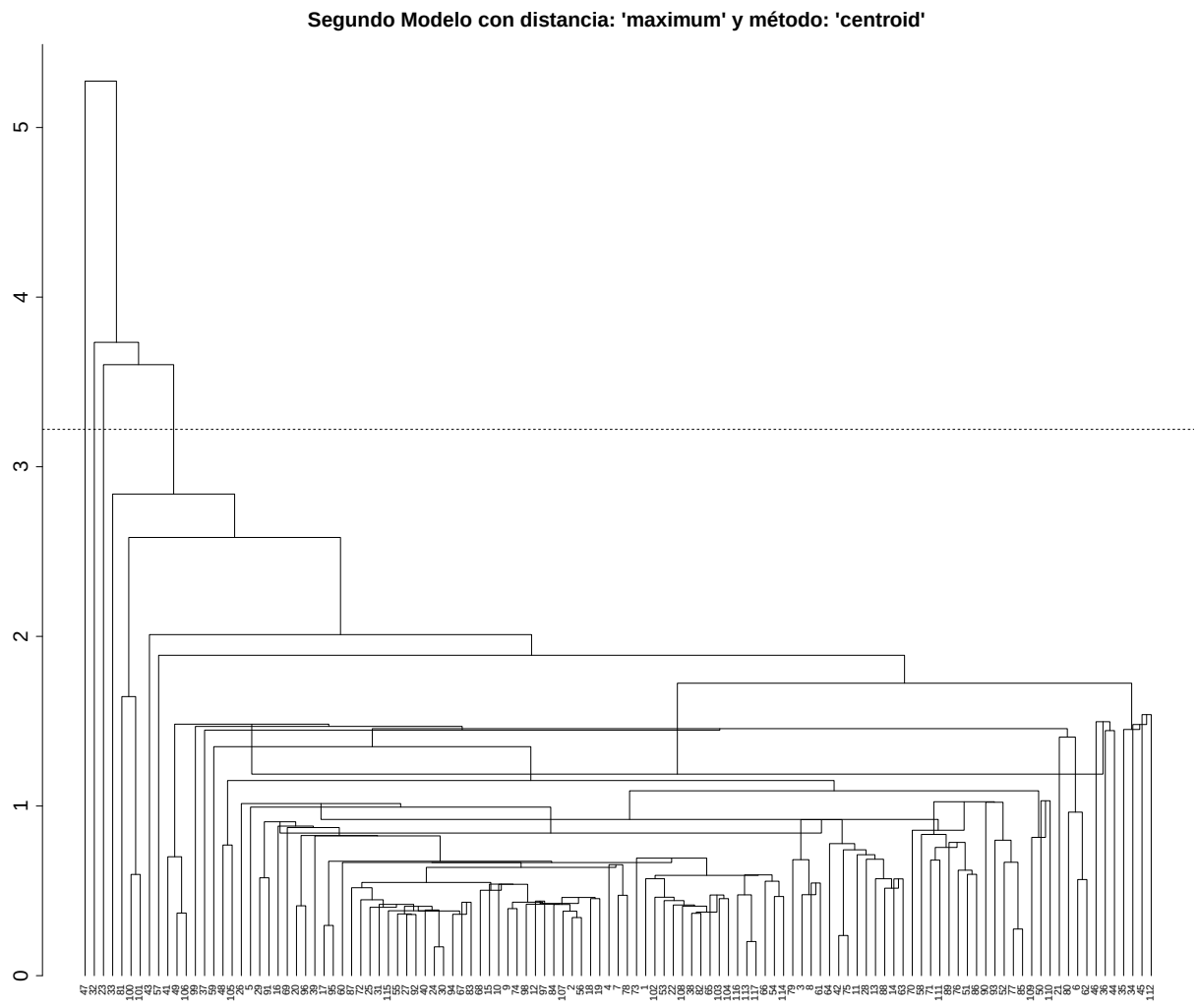
Table 1: Tabla de los distintos modelos -considerando distintas matrices de distancias y linkage intercluster- y sus respectivos coeficientes cofeneticos (orden descendente)

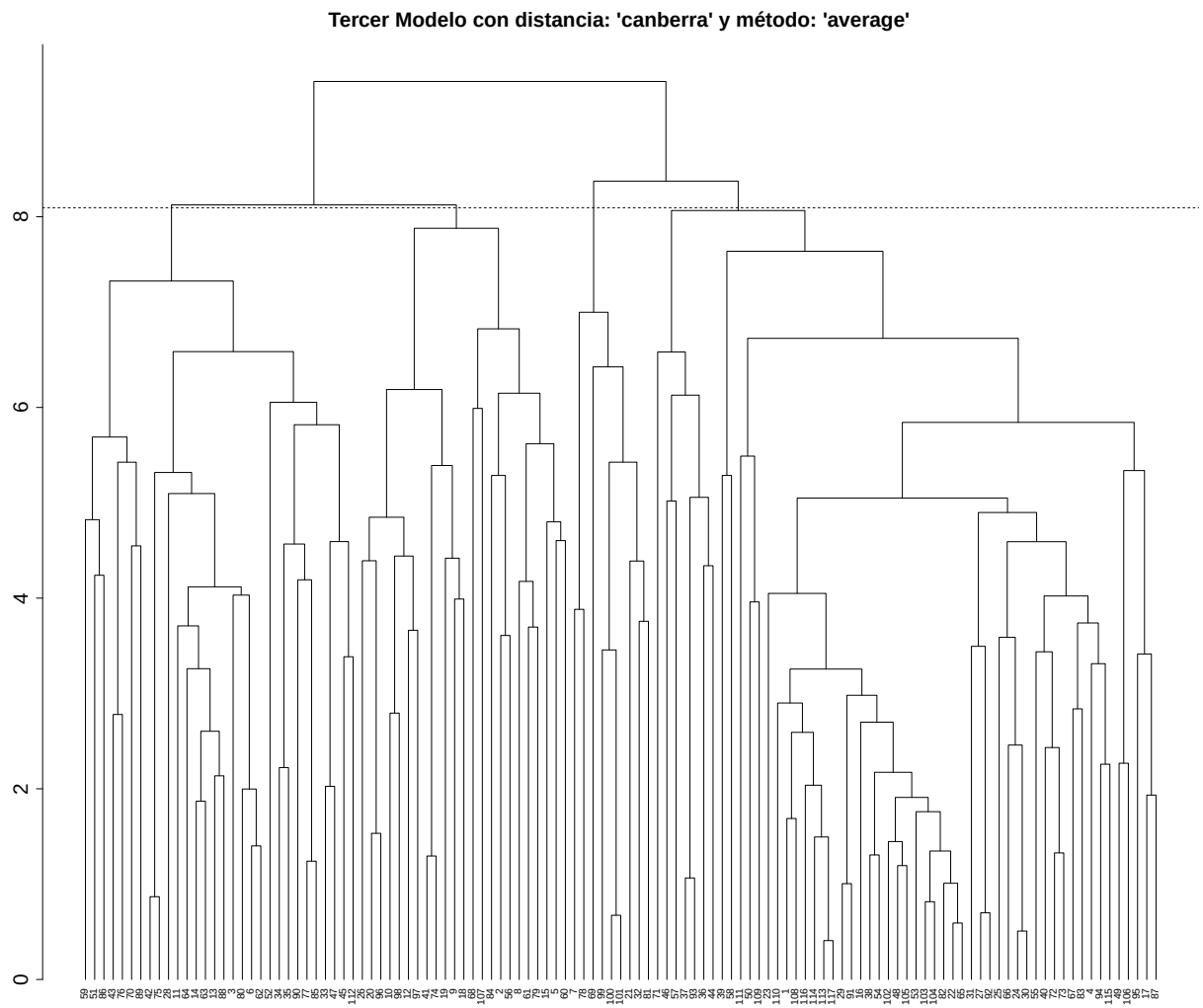
distancias	metodos_linkage	coeficiente_cophenetic
maximum	average	0.8571436
maximum	centroid	0.8394245
canberra	average	0.8240542
canberra	complete	0.7879178
euclidean	average	0.7849688
minkowski	average	0.7849688
euclidean	centroid	0.7564359
minkowski	centroid	0.7564359
manhattan	average	0.7444833
euclidean	single	0.7423286
minkowski	single	0.7423286
canberra	single	0.7165870
manhattan	single	0.6984207
manhattan	centroid	0.6942181

maximum	single	0.6706084
canberra	centroid	0.6618813
maximum	complete	0.6064466
manhattan	complete	0.5835342
canberra	ward	0.5761555
euclidean	complete	0.5211826
minkowski	complete	0.5211826
manhattan	ward	0.4186393
euclidean	ward	0.4002712
minkowski	ward	0.4002712
maximum	ward	0.3733369
binary	complete	NA
binary	average	NA
binary	single	NA
binary	centroid	NA
binary	ward	NA

Modelo Óptimo con distancia: 'maximum' y método: 'average'




Segundo Modelo con distancia: 'maximum' y método: 'centroid'
0
1
2
3
4
5



2.9 Estudio de la tendencia de clustering

```
[1] 0.7515917
```

```
[1] "Los datos presentan agrupamientos importante, con el estadístico Hopkins <= 0.75"
```

3 Pregunta3

3.1 EDA

3.1.1 structure

```
tibble [34 x 3] (S3: tbl_df/tbl/data.frame)
 $ longitud_pata  : num [1:34] 191 185 200 173 171 160 188 186 174 163 ...
 $ circunf_abdomen: num [1:34] 131 134 137 127 128 118 134 129 131 115 ...
 $ long_antena    : num [1:34] 53 50 52 50 49 47 54 51 52 47 ...
```

3.1.2 Summary

```
 longitud_pata   circunf_abdomen  long_antena
 Min.   :160.0   Min.   :107.0   Min.   :43.00
 1st Qu.:176.2   1st Qu.:121.2   1st Qu.:48.25
 Median :187.5   Median :126.0   Median :50.00
 Mean   :193.8   Mean   :125.1   Mean   :49.71
 3rd Qu.:208.8   3rd Qu.:130.5   3rd Qu.:52.00
 Max.   :242.0   Max.   :144.0   Max.   :54.00
```

3.1.3 Control NAs

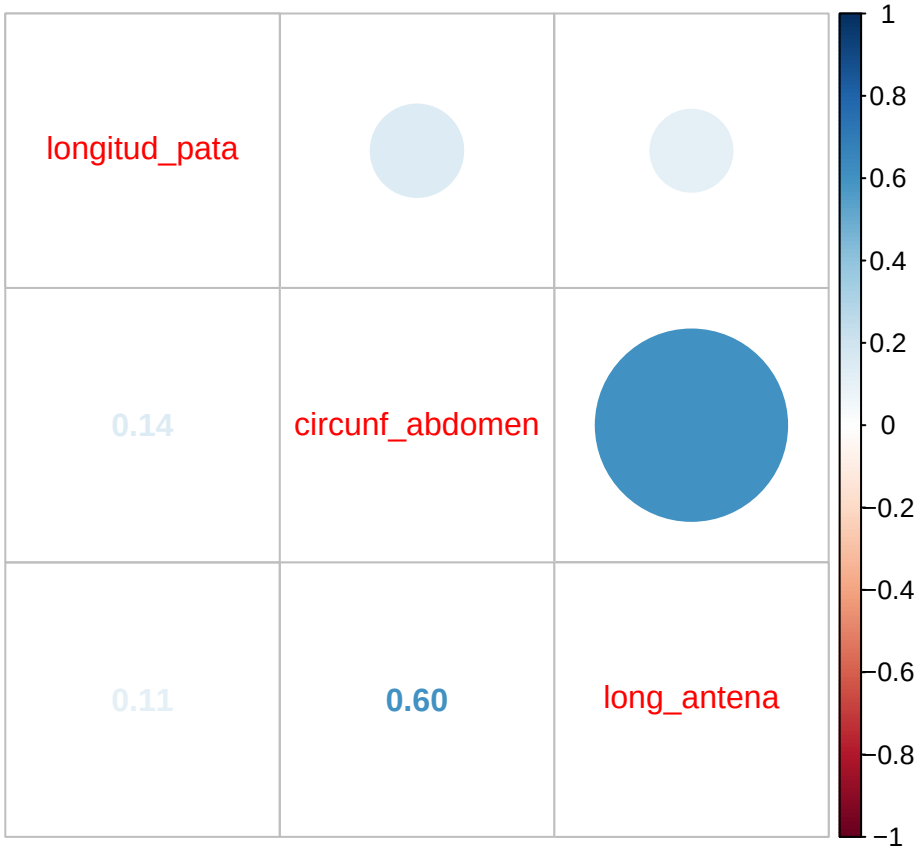
```
# A tibble: 1 x 3
  longitud_pata circunf_abdomen long_antena
          <int>           <int>       <int>
1             0               0           0
```

3.1.4 Distribución de datos

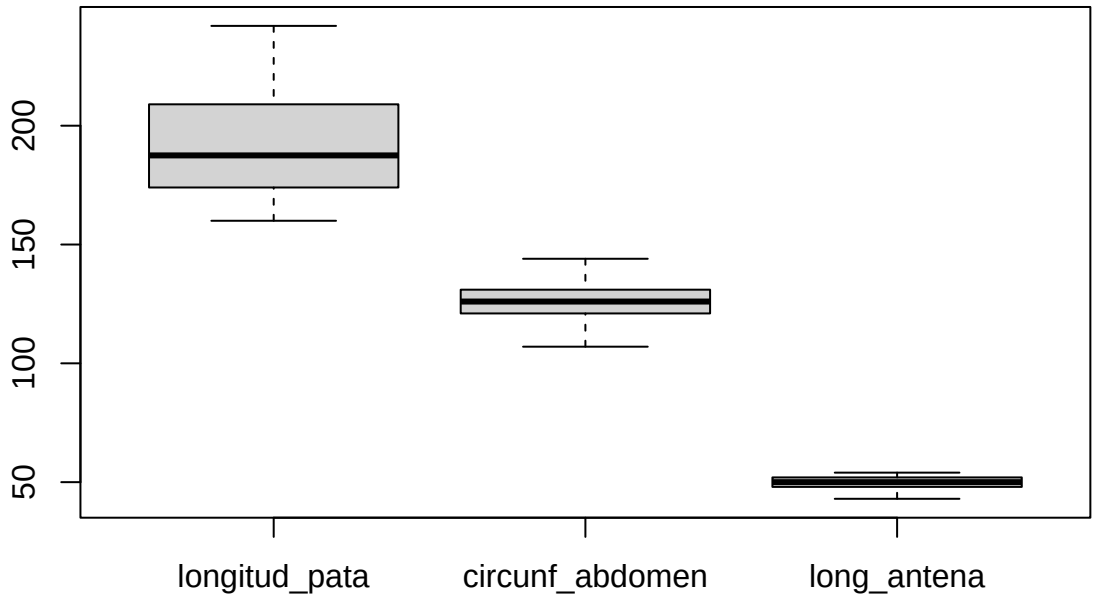
```
$coeficiente_variacion
# A tibble: 1 x 3
  longitud_pata circunf_abdomen long_antena
          <dbl>           <dbl>       <dbl>
1          11.4            6.68        5.35

$sesgo
# A tibble: 1 x 3
  longitud_pata circunf_abdomen long_antena
          <dbl>           <dbl>       <dbl>
1         0.674          -0.307      -0.455

$curtosis
# A tibble: 1 x 3
  longitud_pata circunf_abdomen long_antena
          <dbl>           <dbl>       <dbl>
1          2.83            3.14        3.01

$mad
# A tibble: 1 x 3
  longitud_pata circunf_abdomen long_antena
          <dbl>           <dbl>       <dbl>
1          20.8            7.41        2.97

$m_correlacion
                longitud_pata circunf_abdomen long_antena
```

```
longitud_pata          1.00          0.14          0.11
circunf_abdomen        0.14          1.00          0.60
long_antena            0.11          0.60          1.00
```

3.1.5 Grafico Correlaciones



3.1.6 Boxplot variables numericas



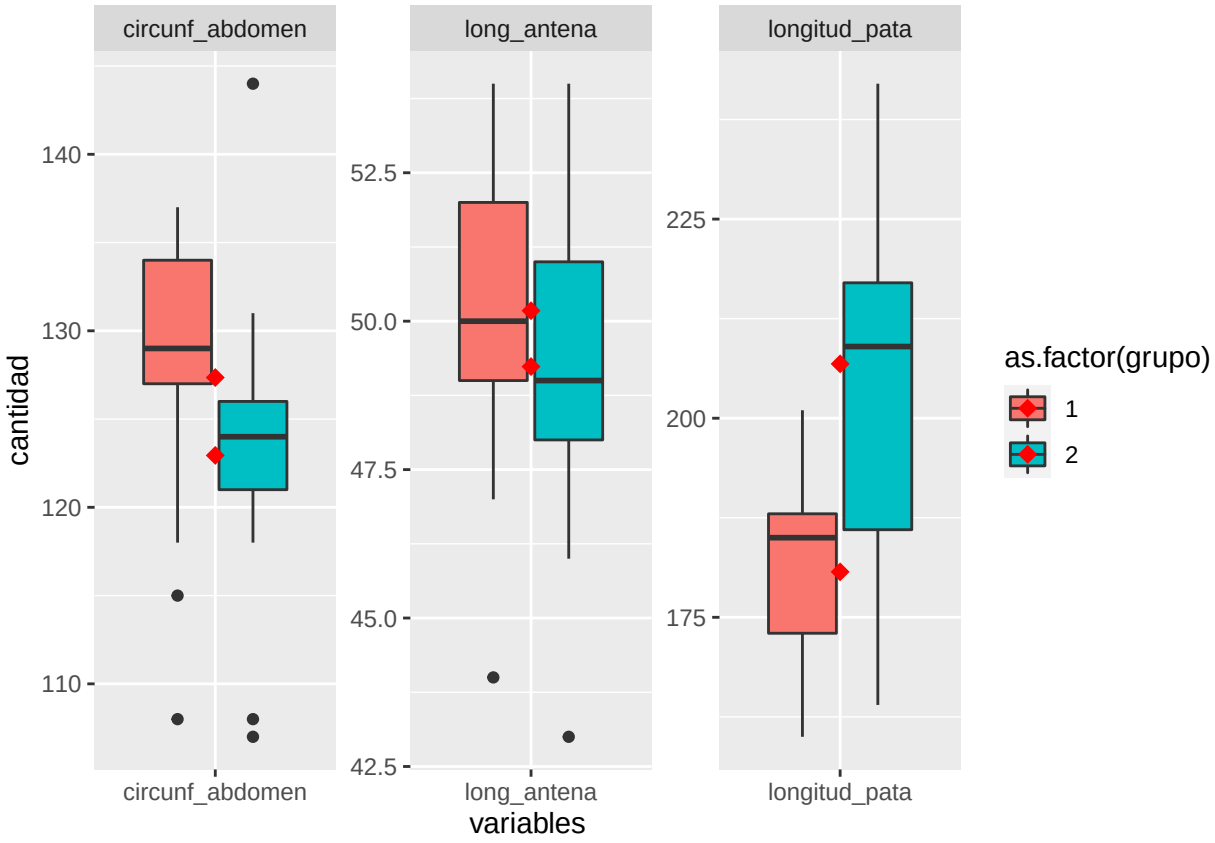
3.1.7 Multigráficos

Considerando la consigna aplicarmos Análisis discriminante lineal.

3.2 Analisis Discriminante Lineal (LDA)

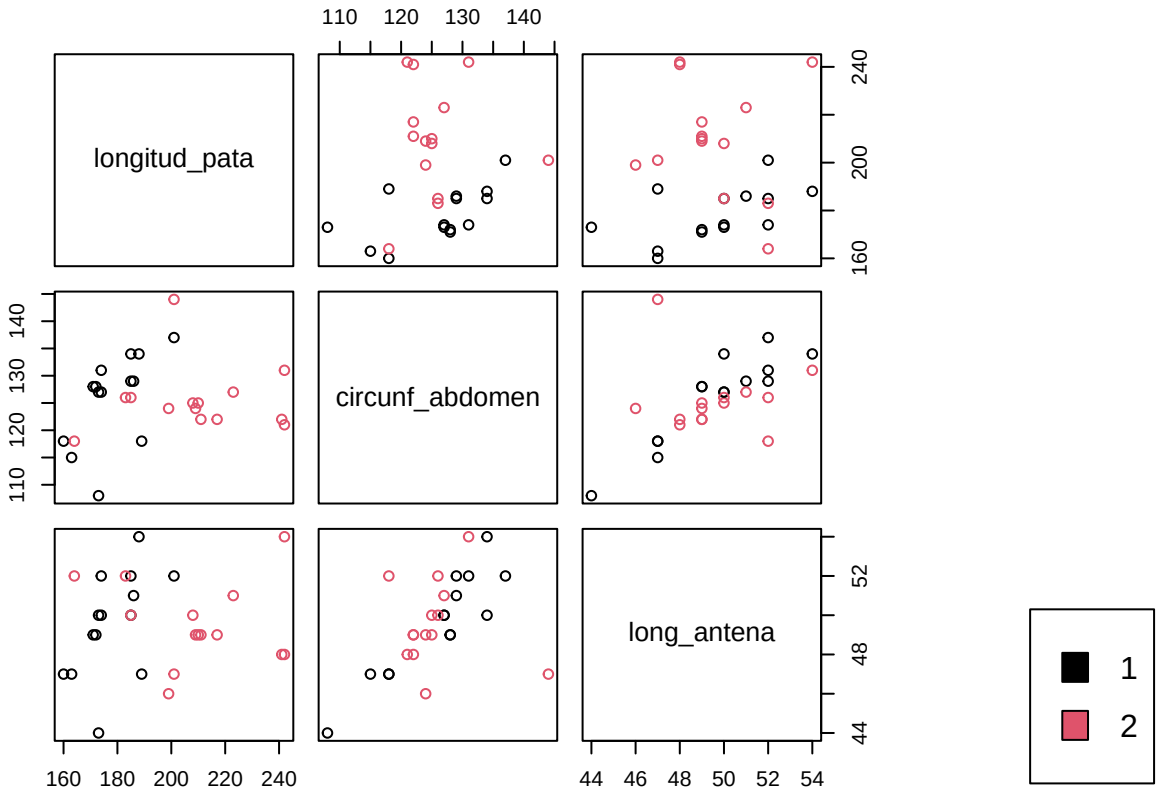
```
tibble [34 x 4] (S3: tbl_df/tbl/data.frame)
 $ grupo          : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
 $ longitud_pata  : num [1:34] 191 185 200 173 171 160 188 186 174 163 ...
 $ circunf_abdomen: num [1:34] 131 134 137 127 128 118 134 129 131 115 ...
 $ long_antena    : num [1:34] 53 50 52 50 49 47 54 51 52 47 ...
```

3.2.1 Box por variable

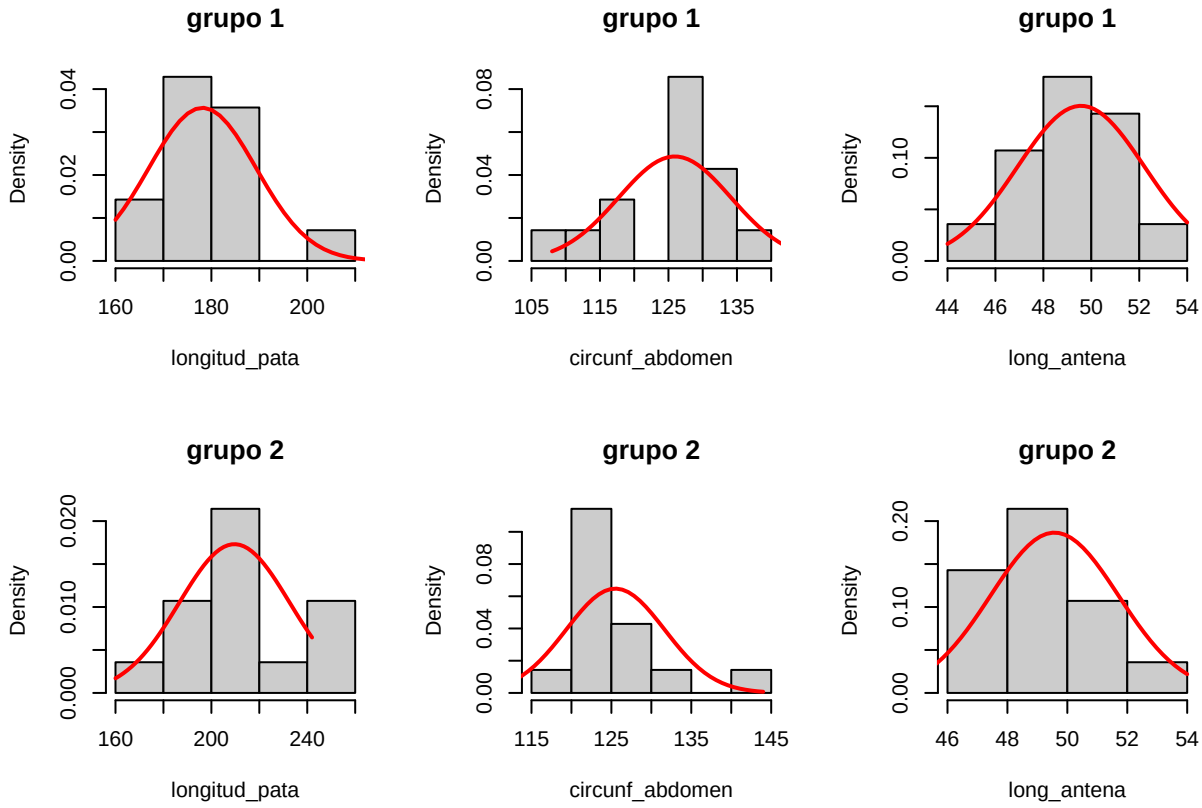


Del gráfico surge que la variable de mayor poder discriminante sería la longitud de la pata en estos insectos.

3.2.2 Explorando discriminación por pares de variable



3.2.3 Histograma VariablexGrupo



3.2.4 Contraste de Normalidad Univariante Shapiro-Wilk

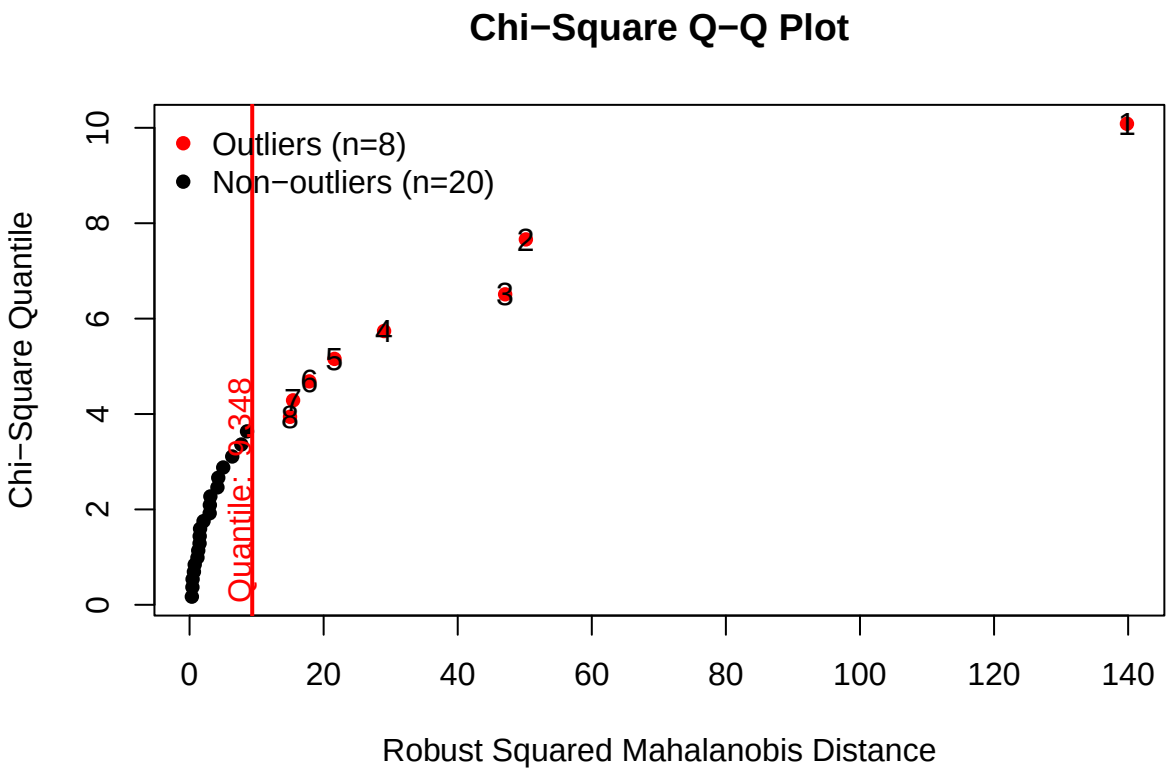
train_tidy[["grupo"]]	variable	p_value_Shapiro.test
1	longitud_pata	0.48753
1	circunf_abdomen	0.15230
1	long_antena	0.69680
2	longitud_pata	0.51305
2	circunf_abdomen	0.00227
2	long_antena	0.83309

```
[1] "H0 debe rechazarse: hay evidencia de falta de normalidad en los siguientes casos"
```

```
# A tibble: 1 x 3
# Groups:   train_tidy[["grupo"]] [1]
  `train_tidy[["grupo"]]` variable        p_value_Shapiro.test
  <fct>                   <fct>                          <dbl>
1 2                       circunf_abdomen              0.00227
```

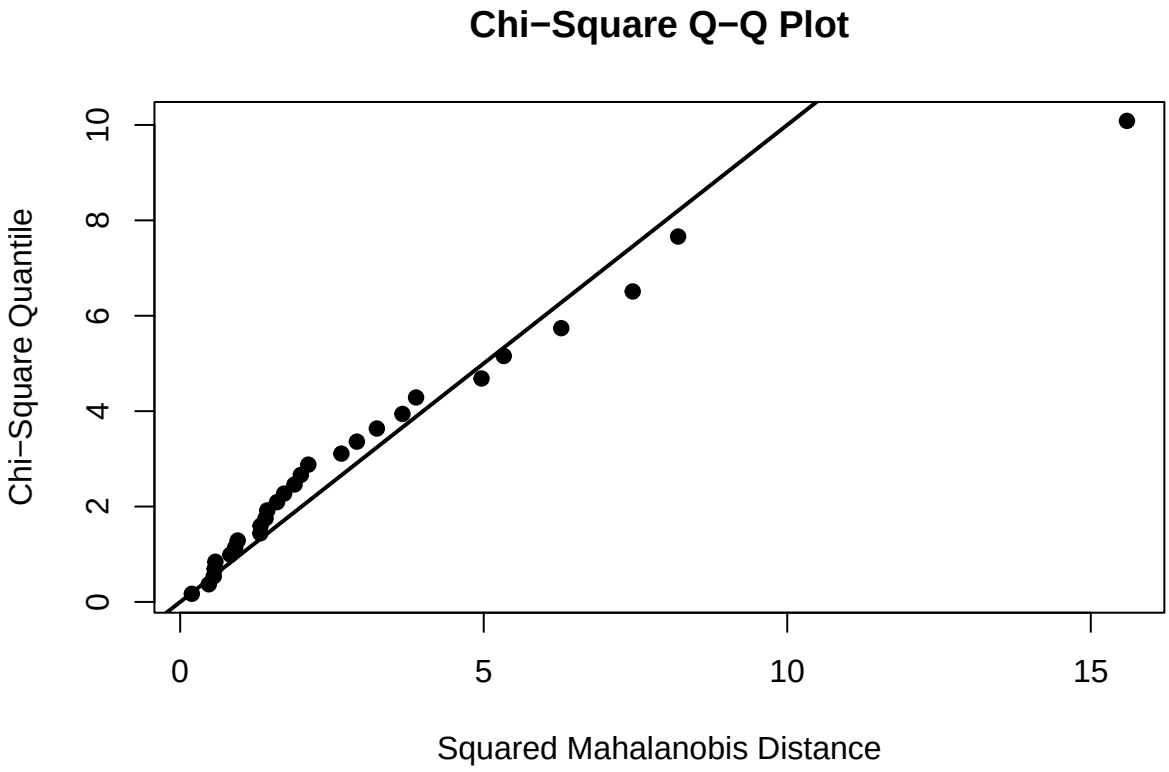
3.2.5 Contraste de Normalidad MultiVariante

3.2.6 Outliers



Vemos la presencia de ouliers

3.2.7 Test de Royston




```
     Test        H   p value MVN
1 Royston 4.841322 0.1868472 YES
```

```
[1] "No hay evidencia de falta de normalidad multivariante a nivel de significancia 0.05"
```

3.2.8 Test de Henze-Zirkler

```
           Test        HZ    p value MVN
1 Henze-Zirkler 0.9552214 0.01986132  NO
```

```
[1] "H0 debe rechazarse: falta de normalidad multivariante a nivel de significancia 0.05"
```

3.2.9 Contraste Homosedasticidad

3.2.10 Test de Levene

Mas robusto que el test M de Box

```
Levene's Tests for Homogeneity of Variance (center = median)

               df1 df2 F value  Pr(>F)
longitud_pata    1  26  3.3224 0.07986 .
circunf_abdomen  1  26  1.0605 0.31260
long_antena      1  26  0.4958 0.48763
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
[1] "No hay evidencia para rechazar H0, luego los datos son homosedásticos"
```

3.2.11 Estimación de parámetros de la función de densidad y cálculo de la función discriminante según aproximación de Fisher via lda()

```
Call:
lda(temp, train[[{
    {
        variable_factor_lda
    }
}]])

Prior probabilities of groups:
  1   2
0.5 0.5

Group means:
  longitud_pata circunf_abdomen long_antena
1      178.1429        125.9286    49.57143
2      209.6429        125.5000    49.57143

Coefficients of linear discriminants:
                         LD1
longitud_pata    0.056647872
circunf_abdomen -0.038191217
long_antena      0.008570581
```

3.2.12 Evaluación del error en Test Set: Accuracy Table

```
          Clase predicha
Clase real 1 2
         1 3 0
         2 0 3
```

3.2.13 Precisión del modelo en test set

```
[1] 100
```

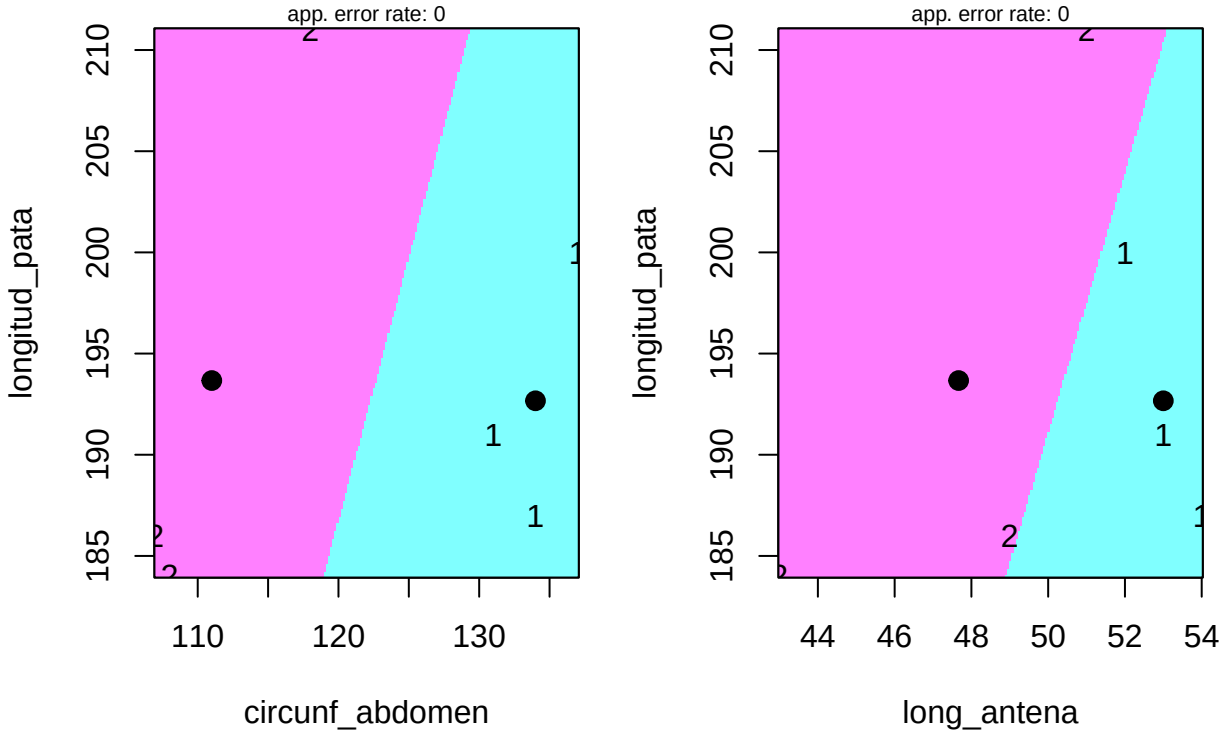
3.2.14 Error en test set

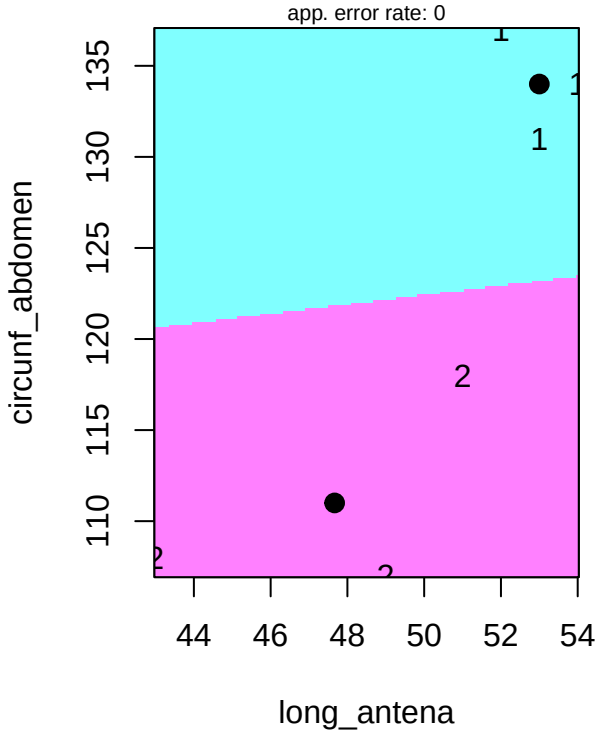
```
[1] "test_error = 0 %"
```

3.2.15 Validación Cruzada (leave one out)

```
[1] 0.25
```

3.2.16 Visualización de las clasificaciones



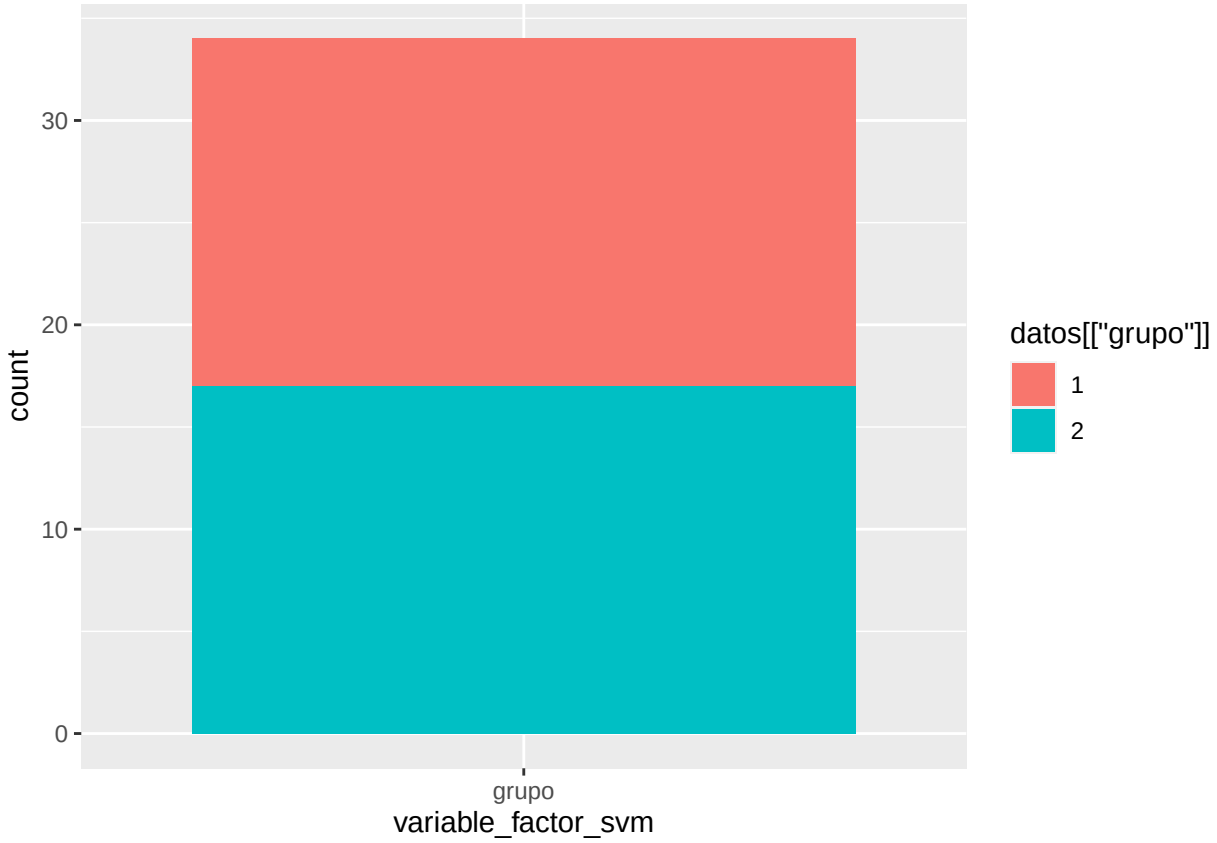


A continuación emplearemos un segundo modelo de clasificación

3.3 Máquinas de Soporte Vectorial

```
tibble [34 x 4] (S3: tbl_df/tbl/data.frame)
 $ grupo          : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
 $ longitud_pata  : num [1:34] 191 185 200 173 171 160 188 186 174 163 ...
 $ circunf_abdomen: num [1:34] 131 134 137 127 128 118 134 129 131 115 ...
 $ long_antena    : num [1:34] 53 50 52 50 49 47 54 51 52 47 ...
```

3.3.1 Grafico datos




```
[1] 28  4
```

```
[1] 6 4
```

3.3.2 Busqueda de mejor hiperparametro C (coste) y Entrenamiento del Modelo con kernel lineal

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost
  0.1

- best performance: 0.2166667

- Detailed performance results:
    cost     error dispersion
1  0.001 0.6333333  0.2459549
2  0.010 0.6333333  0.2459549
3  0.100 0.2166667  0.2490724
4  1.000 0.2166667  0.2490724
5  5.000 0.2166667  0.2490724
6 10.000 0.2166667  0.2490724
7 15.000 0.2166667  0.2490724
```

```
8 20.000 0.2166667  0.2490724
```

3.3.2.1 Mejor modelo según hiperparametro

```
Call:
best.tune(method = svm, train.x = temp, train.y = datos_train[[{
    {
        variable_factor_svm
    }
}]], ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 15, 20)),
    kernel = "linear", scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  0.1

Number of Support Vectors:  25

 ( 13 12 )


Number of Classes:  2

Levels:
 1 2

[1] 1 2 3 4 6 7
```

3.3.3 Predicciones del Modelo

```
          real
prediccion 1 2
         1 3 0
         2 0 3

[1] "Observaciones de test mal clasificadas: 0 %"

[1] "Observaciones de test bien clasificadas: 100 %"
```

3.3.4 Busqueda de mejor hiperparametro C (coste) y Entrenamiento del Modelo con kernel polynomial

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost
   15

- best performance: 0.1833333

- Detailed performance results:
    cost     error dispersion
```

```
1  0.001 0.6833333  0.1229775
2  0.010 0.6833333  0.1229775
3  0.100 0.4833333  0.3884919
4  1.000 0.4500000  0.3518031
5  5.000 0.3000000  0.2810913
6 10.000 0.2166667  0.2944969
7 15.000 0.1833333  0.2539807
8 20.000 0.1833333  0.2539807
```

3.3.4.1 Mejor modelo según hiperparametro

```
Call:
best.tune(method = svm, train.x = temp, train.y = datos_train[[{
    {
        variable_factor_svm
    }
}]], ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 15, 20)),
    kernel = "polynomial", scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  15
     degree:  3
     coef.0:  0

Number of Support Vectors:  17

 ( 9 8 )


Number of Classes:  2

Levels:
 1 2
```

3.3.5 Predicciones del Modelo

```
          real
prediccion 1 2
         1 3 1
         2 0 2
```

```
[1] "Observaciones de test mal clasificadas: 16.67 %"
```

```
[1] "Observaciones de test bien clasificadas: 83.33 %"
```

3.3.6 Busqueda de mejor hiperparametro C (coste) y Entrenamiento del Modelo con kernel sigmoid

3.3.6.1 Mejor modelo según hiperparametro

```
Call:
best.tune(method = svm, train.x = temp, train.y = datos_train[[{
    {
```

```
        variable_factor_svm
    }
}]], ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 15, 20)),
    kernel = "sigmoid", scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  1
     coef.0:  0

Number of Support Vectors:  18

 ( 9 9 )


Number of Classes:  2

Levels:
 1 2
```

3.3.7 Predicciones del Modelo

```
          real
prediccion 1 2
         1 3 0
         2 0 3
```

```
[1] "Observaciones de test mal clasificadas: 0 %"
```

```
[1] "Observaciones de test bien clasificadas: 100 %"
```

3.3.8 Busqueda de mejor hiperparametro C (coste) y Entrenamiento del Modelo con kernel radial

3.3.8.1 Mejor modelo según hiperparametro

```
Call:
best.tune(method = svm, train.x = temp, train.y = datos_train[[{
    {
        variable_factor_svm
    }
}]], ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 15, 20)),
    kernel = "radial", scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  5

Number of Support Vectors:  15

 ( 6 9 )
```

```
Number of Classes:  2

Levels:
 1 2
```

3.3.9 Predicciones del Modelo

```
          real
prediccion 1 2
         1 3 1
         2 0 2
```

```
[1] "Observaciones de test mal clasificadas: 16.67 %"
```

```
[1] "Observaciones de test bien clasificadas: 83.33 %"
```

3.3.10 Respuestas

Considerando las dos metodologías, aunque la precisión de ambos modelos LDA y SVM (con kernel lineal) es la misma (100% de precisión en test set), consierando la poca cantidad de datos elegirìa el SVM.

4 Documento e Información de Sesion

Este documento fue generado a partir de un documento **RMarkdown** parametrizable que se entrega con su correspondiente output en pdf. Para conrroborar la originalidad y autoría del documento .Rmd se pone a disposición (bajo requerimiento) de la fecha de ceación e hitórico de cambios alojado en la cuenta personal del suscripto castillosebastian@github.com.

```
R version 4.2.0 (2022-04-22)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.4 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/liblapack.so.3

locale:
 [1] LC_CTYPE=es_AR.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=es_AR.UTF-8        LC_COLLATE=es_AR.UTF-8
 [5] LC_MONETARY=es_AR.UTF-8    LC_MESSAGES=es_AR.UTF-8
 [7] LC_PAPER=es_AR.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=es_AR.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] grid      stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] dplyr_1.0.9        raster_3.5-15      sp_1.4-7           cluster_2.1.3
 [5] readxl_1.4.0       caret_6.0-92       lattice_0.20-45    e1071_1.7-9
 [9] biotools_4.2       klaR_1.7-0         MVN_5.9            reshape2_1.4.4
[13] nortest_1.0-4      gplots_3.1.3       vcd_1.4-9          factoextra_1.0.7
[17] FactoMineR_2.4     broom_0.8.0        MASS_7.3-56        htmltools_0.5.2
[21] moments_0.14.1     corrplot_0.92      skimr_2.1.4        jsonlite_1.8.0
[25] formattable_0.2.1  tibbletime_0.1.6   readr_2.1.2        ggthemes_4.2.4
[29] rlang_1.0.2        gghighlight_0.3.2  scales_1.2.0       lubridate_1.8.0
[33] colorRamps_2.3.1   RColorBrewer_1.1-3 ggbeeswarm_0.6.0   ggplot2_3.3.6
[37] tibble_3.1.7       tidyr_1.2.0        kableExtra_1.3.4   janitor_2.1.0
[41] stringr_1.4.0      knitr_1.39

loaded via a namespace (and not attached):
  [1] utf8_1.2.2           questionr_0.7.7      tidyselect_1.1.2
  [4] htmlwidgets_1.5.4    combinat_0.0-8       pROC_1.18.0
  [7] munsell_0.5.0        codetools_0.2-18     DT_0.22
 [10] future_1.25.0        miniUI_0.1.1.1       withr_2.5.0
 [13] colorspace_2.0-3     energy_1.7-10        highr_0.9
 [16] rstudioapi_0.13      leaps_3.1            stats4_4.2.0
 [19] ggsignif_0.6.3       listenv_0.8.0        labeling_0.4.2
 [22] repr_1.1.4           mnormt_2.0.2         bit64_4.0.5
 [25] farver_2.1.0         parallelly_1.31.1    vctrs_0.4.1
 [28] generics_0.1.2       ipred_0.9-12         xfun_0.30
 [31] R6_2.5.1             bitops_1.0-7         reshape_0.8.9
 [34] promises_1.2.0.1     vroom_1.5.7          nnet_7.3-17
 [37] beeswarm_0.4.0       gtable_0.3.0         globals_0.15.0
 [40] timeDate_3043.102    systemfonts_1.0.4    scatterplot3d_0.3-41
```

```
 [43] splines_4.2.0        rstatix_0.7.0        ModelMetrics_1.2.2.2
 [46] yaml_2.3.5           abind_1.4-5          backports_1.4.1
 [49] httpuv_1.6.5         tools_4.2.0          lava_1.6.10
 [52] psych_2.2.3          ellipsis_0.3.2       proxy_0.4-26
 [55] Rcpp_1.0.8.3         plyr_1.8.7           base64enc_0.1-3
 [58] purrr_0.3.4          ggpubr_0.4.0         rpart_4.1.16
 [61] zoo_1.8-10           haven_2.5.0          ggrepel_0.9.1
 [64] magrittr_2.0.3       data.table_1.14.2    lmtest_0.9-40
 [67] tmvnsim_1.0-2        gsl_2.1-7.1          hms_1.1.1
 [70] mime_0.12            evaluate_0.15        xtable_1.8-4
 [73] compiler_4.2.0       KernSmooth_2.23-20   crayon_1.5.1
 [76] later_1.3.0          tzdb_0.3.0           boot_1.3-28
 [79] Matrix_1.4-1         car_3.0-13           cli_3.3.0
 [82] heplots_1.3-9        parallel_4.2.0       gower_1.0.0
 [85] forcats_0.5.1        pkgconfig_2.0.3      flashClust_1.01-2
 [88] terra_1.5-21         recipes_0.2.0        xml2_1.3.3
 [91] foreach_1.5.2        svglite_2.1.0        vipor_0.4.5
 [94] hardhat_0.2.0        webshot_0.5.3        prodlim_2019.11.13
 [97] rvest_1.0.2          snakecase_0.11.0     digest_0.6.29
[100] rmarkdown_2.14       cellranger_1.1.0     shiny_1.7.1
[103] gtools_3.9.2         lifecycle_1.0.1      nlme_3.1-157
[106] carData_3.0-5        viridisLite_0.4.0    fansi_1.0.3
[109] labelled_2.9.1       pillar_1.7.0         GGally_2.1.2
[112] fastmap_1.1.0        httr_1.4.3           survival_3.3-1
[115] glue_1.6.2           iterators_1.0.14     bit_4.0.4
[118] class_7.3-20         stringi_1.7.6        caTools_1.18.2
[121] future.apply_1.9.0
```